Calculating Standard Errors of Predicted Values Based on Nonlinear Functional Forms

But suppose we consider what mechanism might drive heteroskedasticity, not as a statistical annoyance but as a substantively important phenomenon. For example, candidates are said to relish ambiguity. But what effect does ambiguity have on voter perceptions of candidate positions? Surely not through changing the (conditional) mean of perception. Rather ambiguity shows itself in the variance surrounding the (conditional) mean. In short, variation in ambiguity implies heteroskedasticity. Our substantive concern is how ambiguity affects voter uncertainty about candidate location and our statistical model must incorporate this substantive hypothesis. For a ML based solution to this problem, see Franklin, *APSR*, Dec. 1991.

While we can "cure" heteroskedasticity using least squares just as well as we can using ML, the latter offers us greater flexibility to look at heteroskedasticity in a new light. That is we can view it as something we want to model because it incorporates substantive meaning.

How should we model heteroskedasticity? ML methods have the great advantage of making it easy to do, rather than making it possible to do. We could do something like this with least squares, it simply turns out to be easier to do using ML. And we can then get good tests of our substantive results. This is an irresistible combination.

As with any statistical method, ML is no panacea. As with regression, structural equations and LISREL before it, ML will take time for us to come to grips with its advantages and its drawbacks. More importantly, we need to educate ourselves and our students in the proper use of this approach to estimation. It is important to realize that ML is a method of estimation, not a particular statistical model. Thus there are many models, including simple regressions, which can be estimated using ML. It is not the method of estimation which is important. It is the flexibility we gain in model specification and the ease of estimation which ML provides which makes it worth learning.

This raises the pedagogical issues that we need to face. If political methodology is to serve to advance the substantive work of the discipline, then it must be increasingly tied to the solutions of the kinds of substantive problems we face. ML is a very flexible tool that lets us do that better than we can with other approaches. This is its strength, and this is why we should learn about it and teach about it. How we go about doing that is a topic for another issue of *TPM*.

I am anxious to receive articles on the direction of methods requirements and offerings in departments around the country. We face a significant problem of teaching increasingly advanced techniques to graduate students who are as unprepared as ever. Articles addressing this subject would be welcome. The deadline for the next issue of *TPM* is March 15, 1992. Correspondence related to that issue should be sent to me at Washington University. —*CHF*

Calculating Standard Errors of Predicted Values based on Nonlinear Functional Forms

Gary King[1]
Harvard University

Introduction

Whenever we report predicted values, we should also report some measure of the uncertainty of these estimates. In the linear case, this is relatively simple, and the answer well-known, but with nonlinear models the answer may not be apparent. This short article shows how to make these calculations. I first present this for the familiar linear case, also reviewing the two forms of uncertainty in these estimates, and then show how to calculate these for any arbitrary function. An example appears last.

The Linear Case

Suppose we run a linear regression of y on X producing a vector of coefficients, $b = (X'X)^{-1}X'y$. Now think about setting the explanatory variables to new values (X^p) and making a prediction under this new situation about the next value of the random variable Y, which we will call Y^p. Y^p has N elements, which is the number of predicted values (not the number of observations).

The prediction Y^p has two forms of uncertainty associated with it: The first is that its expected value $E(Y^p)$ is estimated (with $\hat{y}^p = X^p b$) and thus depends on the estimated value b. This *estimation uncertainty* can be systematically reduced by increasing the sample size. The second is *fundamental variability* in the dependent variable around this expected value. Thus, even if we knew β (as in the case where we had an infinite number of observations), and were thus able to observe $E(Y^p) = \mu = X^p\beta$, we would not expect that μ would provide perfect predictions.

In the linear case, we can think about the components of the variance in Y^p simply by writing down the prediction as $Y^p = X^p b + \epsilon^p$ and calculating its variance:

$$\begin{aligned} V(Y^p) &= V(X^p b) + V(\epsilon^p) \\ &= X^p V(b) X^{p\prime} + \sigma^2 I \\ &= \sigma^2 X^p (X^{p\prime} X^p)^{-1} X^{p\prime} + \sigma^2 I \end{aligned} \tag{1}$$

where the two terms on the right side of this equation correspond to the estimation uncertainty and the fundamental variability, respectively. This is a simpler form of the proof in Johnston (1984: 199).

[1]Thanks to Neal Beck and Andrew Gelman.

To make the transition from the linear case to the more general case, we write this down and then analyze the full likelihood model (although in the linear case, one can make less restrictive assumptions). Thus, we assume a Normal stochastic component:

$$Y \sim N(\mu, \sigma^2 I) \tag{2}$$

and linear systematic component, $\mu = X\beta$. The likelihood function is then:

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{n} N(X_i\beta, \sigma^2) \tag{3}$$

Under a properly specified likelihood model, such as this, the maximum likelihood estimate has an asymptotic Normal sampling distribution (see King, 1989, for a review of likelihood models and methods). In this case, the estimator b of β is distributed as follows:

$$\begin{aligned} b &\sim N(\beta, V(b)) \\ &= N\left(\beta, \sigma^2(X'X)^{-1}\right) \end{aligned} \tag{4}$$

We can now specify the distribution of the two quantities of interest in this analysis, $\hat{Y}^p$ and Y^p.

$$\hat{Y}^p \sim N(X^p\beta,\ X^pV(b)X^{p\prime}) \tag{5}$$

and

$$Y^p|\beta \sim N(X^p\beta,\ \sigma^2 I) \tag{6}$$

Since these distributions, representing estimation variability and fundamental variability respectively, are independent, we can combine them to derive the unconditional distribution for Y^p.

$$Y^p \sim N(X^p\beta,\ X^pV(b)X^{p\prime} + \sigma^2 I) \tag{7}$$

If we are really interested in knowing where the expected value is, then we would focus on the estimation variability in Equation 5. At other times we might be interested only in the variability around the expected value—variability in the world which beyond that due to estimation; this can be found in the fundamental variability of Equation 6. However, in most cases, we are interested in making predictions of where Y^p is likely to be on the basis of some specified explanatory variables X^p, as can be found in the total variability represented in Equation 7.

All Other Cases

To begin, I write down a general likelihood specification, with a stochastic component

$$Y \sim f(y|\theta) \tag{8}$$

and a (linear or nonlinear) systematic component, which for simplicity we assume is the expected value of the random variable:

$$E(Y) = \theta = g(X, \beta) = g(\beta) \tag{9}$$

The likelihood function is then written as usual

$$L(\beta|y) = \prod_{i=1}^{n} f(y_i|\theta) \tag{10}$$

The estimate of $E(Y)$ is the predicted value

$$\hat{y}^p = g(X^p, b) = g(b) \tag{11}$$

where b is the ML estimator of β. As in the linear case, the explanatory variable matrix X^p has one row for each predicted value, and so has dimensions $(N \times k)$.

Some examples of nonlinear functional forms include the exponential, logistic, or probit functions, respectively:

$$E(Y_1) = \exp(X\beta) \tag{12}$$

$$E(Y_2) = \frac{1}{1 + \exp(-X\beta)} \tag{13}$$

$$E(Y_3) = \Phi^{-1}(X\beta) \tag{14}$$

The fundamental variability can be calculated, conditional on knowing β, by the usual methods from probability theory:

$$V(Y) = E(Y^2) - E(Y)^2 \tag{15}$$

where

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y|\theta)dy \tag{16}$$

and

$$E(Y) = \int_{-\infty}^{\infty} yf(y|\theta)dy \tag{17}$$

In practice, one usually need not use Equations 15, 16, and 17 since the results for most popular distributions are widely available in books on probability theory. Thus, the fundamental variability is λ in the Poisson distribution, $\lambda\sigma^2$ for the negative binomial (depending on parameterization), and σ^2 for the Normal.

Calculating the estimation variability will usually require more effort, since these calculations are not as widely reported. Since expectations and variances are linear operators, a general method of calculating these is by calculating the linear approximation to the arbitrary linear function—the Taylor series. The Taylor series approximation of $\hat{y}^p = g(b)$ is as follows:

$$\hat{y}^p = g(b) = g(\beta) + g'(\beta)(b - \beta) + \cdots \tag{18}$$

where $g'(\beta)$ is the first derivative of the functional form $g(\beta)$ (from Equation 9) with respect to β. If there are k elements of β, and N predicted values, $g(\beta)$ is $(N \times 1)$ and $g'(\beta)$ is $(N \times k)$.

We now drop all but the first two terms in Equation 18 (making the equality in that equation an approximation), and apply the variance operator:

$$\begin{aligned} V(\hat{Y}^p) &\approx V[g(\beta)] + V[g'(\beta)(b-\beta)] \\ &= g'(\beta)V(b)g'(\beta)' \end{aligned} \tag{19}$$

(No covariances are necessary because each of the terms in the Taylor series approximation are independent.)

The matrix on the right side of Equation 19 is an $(N \times N)$ matrix, whereas $V(b)$ is $(k \times k)$. The $V(\hat{Y}^p)$ can be consistently estimated by substituting the estimated parameter vector and covariance matrix calculated by all standard ML routines for β and $V(b)$, respectively, in this equation. The standard errors of the elements of $\hat{Y}^p$ (based on estimation variability only) are the square roots of the diagonal elements of this matrix. Off diagonal elements of this matrix are covariances, useful for calculating the variances of constructions such as $(\hat{Y}_1^p - \hat{Y}_2^p)$.[2]

The total variability is again merely the sum of the estimation and fundamental variabilities.

An Example

Consider the Poisson regression model (King, 1989: Chapter 5). The stochastic component is Poisson:

$$Y \sim \frac{e^{-\lambda}\lambda^y}{y!} \tag{20}$$

and the systematic component is exponential:

$$\lambda = \exp(X\beta) \tag{21}$$

The likelihood function is then

$$L(\beta|y) = \prod_{i=1}^{n} \frac{\exp(-e^{X\beta})\exp(X\beta)^y}{y!} \tag{22}$$

The fundamental variability in the Poisson model happens to equal the expected value, so that

$$\begin{aligned} V(Y|\lambda) &= \sum_{y=1}^{\infty} y^2 \frac{e^{-\lambda}\lambda^y}{y!} - \left(\sum_{y=1}^{\infty} y \frac{e^{-\lambda}\lambda^y}{y!}\right)^2 \\ &= (\lambda + \lambda^2) - \lambda^2 \\ &= \lambda \end{aligned} \tag{23}$$

[2] An interesting side point can be studied by extending the Taylor series approximation to a third term, $\frac{1}{2}(b-\beta)'g''(\beta)(b-\beta)$, and calculating $E(\hat{Y}^p)$. The expected value of the second term is zero, and the entire approximation now becomes $E(\hat{Y}^p) \approx g(\beta) + \frac{1}{2}A'V(b)A$, where $A'A = g''(\beta)$, the matrix of second derivatives. (A can be calculated by taking the Cholesky decomposition of the second derivative matrix (in Gauss, for example: `A=CHOL(D)`, where `D` is the second derivative.) This calculation shows the details of the familiar result that $g(E(T)) \neq E(g(T))$, for nonlinear functional forms $g(\cdot)$. However, since the term $\frac{1}{2}A'V(b)A$ goes to zero as the sample size increases, $g(\beta)$ is a reasonable approximation to $E(g(b)) \equiv E(\hat{Y}^p)$ in large samples.

Where the summation is used in place of integration since the distribution is discrete. The vector λ is $(N \times 1)$. By the usual no autocorrelation assumptions of the Poisson regression model, the covariances (based on fundamental variability) between different predicted values are zero. Thus, one can write the $(N \times N)$ fundamental variance matrix as λI, where I is an $(N \times N)$ identity matrix. To estimate this fundamental variability, one would use the $(N \times N)$ matrix $\exp(X^p b)I$.

To calculate the estimation variability, we need the first derivative matrix

$$\begin{aligned} g'(\beta) &= \frac{\partial \exp(X\beta)}{\partial \beta} \\ &= X \cdot e^{X\beta} \end{aligned} \tag{24}$$

Note that this is a slight abuse of standard mathematical notation, used in order to make the transition to computation easier. X is $(N \times k)$ and $e^{X\beta}$ is $(N \times 1)$. The notation "$\cdot$" in this equation refers to multiplying each column of X by $e^{X\beta}$ element by element.[3]

Thus, the estimated variance matrix of the k-vector of predicted values $\hat{Y}^p$ is then as follows:

$$V(\hat{Y}^p) = (X^p \cdot e^{X^p b})\widehat{V(b)}(X^p \cdot e^{X^p b})' \tag{25}$$

One would add $\hat{\lambda}$ or, equivalently, $\exp(X^p b)I$ to the diagonal elements of this matrix in order to calculate the total variability.

Concluding Remarks

An important way to evaluate a statistical model is to calculate the total variability and see whether roughly two-thirds of the observed points are within plus or minus a standard error of the predicted values, and whether about 95% of the observations are within plus or minus two standard errors. Without a calculation such as this, all the standard errors and statistical tests based on this model would be in doubt.

Reference

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* New York: Cambridge University Press.

Johnston, J. 1984. *Econometric Methods.* 3rd edn. New York: McGraw Hill.

[3] In Gauss, one would use `X.*exp(X*b)`.